6
Problems of transfer and interlingua

The last chapter was concerned primarily with problems of analysis, and the main focus was monolingual difficulties arising from the source language text. The next chapter will be concerned with the target language text generation. The present chapter concentrates on the interface between these two monolingual components.

Before discussing the different approaches according to the basic strategies described in Chapter 4 (direct, transfer, interlingua), we shall illustrate a few of the lexical and structural differences between languages.

6.1 Lexical differences

In the last chapter we discussed monolingual ambiguities which cause problems during analysis. **Transfer ambiguities** (or translational ambiguities) arise when a single source language word can potentially be translated by a number of different target language words or expressions, not because the source language word itself is ambiguous but because it is 'ambiguous' from the perspective of another language. Most differences in the lexical systems of languages arise from conceptual differences, but there are also those arising from stylistic or grammatical differences.

Stylistic translational ambiguities occur when the choice of target language lexical equivalent depends on differences of register or text-type. An example is the

French word *domicile* which could appear in English as either *home* or *domicile* according to the type of document being translated. This kind of difference is more common between languages which, unlike English and French, do not share similar cultural backgrounds. Japanese, for instance, has different words for kinship relations depending on whether it is the speaker's or hearer's kin being referred to, e.g. *kanai* '(my) wife' but *okusan* '(your) wife'.

Grammatical translational ambiguities are rather less common. These occur when there is a lexical choice in the target language which is conditioned by grammatical context. A familiar example is the translation of English *know* into French or German, where the choice between *connaître* or *savoir* and between *kennen* or *wissen* depends (roughly) on whether the direct object is a noun-phrase (1) or a subordinate clause or an infinitive (2).

(1a) I know the right answer.
Je connais la bonne réponse.
Ich kenne die richtige Antwort.

(1b) I know the author of that book.
Je connais l'auteur de ce livre.
Ich kenne den Verfasser dieses Buchs.

(2a) I know what the right answer is.
Je sais quelle est la bonne réponse.
Ich weiß, was die richtige Antwort ist.

(2b) I know who the author of that book is
Je sais qui est l'auteur de ce livre.
Ich weiß, wer der Verfasser dieses Buchs ist.

In some cases the choice may be open because both types of structure are possible in the target text (3).

(3) I know the quickest way to get from Norwich to Manchester.
Je connais la route la plus rapide de Norwich à Manchester.
Je sais comment aller le plus vite possible de Norwich à Manchester.
Ich kenne den schnellsten Weg von Norwich nach Manchester.
Ich weiß, wie man am schnellsten von Norwich nach Manchester kommt.

It may be argued that the difference between *connaître/kennen* and *savoir/wissen* corresponds to a conceptual difference between types of knowledge — knowing a fact and knowing how to do something — which is not reflected in English by a lexical choice. This would make the difference 'conceptual' rather than grammatical, though it would still be true that in translating from English into French and German the lexical choice can be made largely on the basis of grammatical facts alone.

A special type of grammatical ambiguity is that concerning anaphora. If the antecedent of a pronoun such as *it* in the sentences (28) of Chapter 5 has not been identified, then it may well have to be done during transfer, unless there is the relatively rare possibility of a straight lexical substitution. The latter is more likely with plural pronouns since a number of 'major' languages make no distinctions of gender (English *they*, German *sie*, Russian *oni*).

Conceptual translational ambiguities are the cause of the greatest problems in translation, and are the principal focus of debates about MT methodologies and system designs. They arise when a single 'concept' represented by one word in one language corresponds to a number of concepts, and hence words, in another language. This is by no means a rare phenomenon, even between closely related languages, and examples are plentiful, as illustrated in (4).

(4a)	English *wall*	German	*Wand* (inside a building)
			Mauer (outside)
(4b)	English *river*	French	*rivière* (general term)
			fleuve (major river, flowing into sea)
(4c)	English *leg*	Spanish	*pierna* (human)
			pata (animal, table)
			pie (chair)
			etapa (of a journey)
		French	*jambe* (human)
			patte (animal, insect)
			pied (table, chair)
			étape (journey)
(4d)	English *blue*	Russian	*goluboi* (pale blue)
			sinii (dark blue)
(4e)	French *louer*	English	*hire* or *rent*
(4f)	French *colombe* German *Taube*	English	*pigeon* or *dove*
(4g)	German *leihen*	English	*borrow* or *lend*
(4h)	English *rice*	Malay	*padi* (unharvested grain)
			beras (uncooked)
			nasi (cooked)
			emping (mashed)
			pulut (glutinous)
			bubor (cooked as a gruel)
(4i)	English *wear*	Japanese	*kiru* (generic)
			haoru (coat or jacket)
			haku (shoes or trousers)
			kaburu (hat)
			hameru (ring or gloves)
			shimeru (belt or tie or scarf)
			tsukeru (brooch or clip)
			kakeru (glasses or necklace)

It should be noted that even these correspondences are simplifications. In (4c), for example, French *pied* corresponds also to English *foot* (human, mountain), and *patte* to *paw* (animal) or *foot* (bird); in (4g) English *borrow* can also be German *borgen*, and *lend* can also be *leisten* ('lend help'). The example in (4i) is complicated by the fact that the Japanese translations of *wear* also translate the English *put on*: whereas English distinguishes the process from the resulting state without regard to the type of clothing, in Japanese it is the other way round. So *kaburu* is used not only for wearing a hat but also for putting it on.

There are many other examples of conceptual differences which could have been given: reflections of environmental or cultural differences, weather terms around the world, names of flora and fauna, culinary terms and so on. One example of this phenomenon which is often cited is the supposed abundance in Eskimo of words for 'snow'. In fact this is one of the great myths of linguistics. Eskimo actually has only two words for 'snow' (*qanik* for 'snow in the air', and *aput* for 'snow on the ground'), which is rather fewer than English, for example, (*snow, sleet, slush, hail, freezing rain, blizzard*).

For translation, there are particular problems caused by what are often referred to as **lexical gaps**, the absence in a language of one single word corresponding to a lexical item in the other language. These are not words specific to a particular culture, such as *mansion, cottage, marmelade, dacha, whiskey, vodka* (in these cases the lexical items are often borrowed untranslated), nor words which in one language carry a range of meanings covered by a variety of words in another as illustrated in (4), but 'universal' concepts which happen not to correspond to a single lexical item in a particular language. For example, English has nothing corresponding to French *gratiner* ('to cook with a cheese coating'), Russian *protalina* ('a place where the snow has melted'), Japanese *ōtosanrin* ('a three-wheeled pickup truck or motor van'). Likewise, English has lexical items which must be translated periphrastically: *snub* in French as something like *infliger un affront*, and in German as perhaps *verächtlich behandlen* or *derb zurückweisen*. In some cases the gaps are specific to grammatical categories; for example, German has adjectival forms for a number of adverbial and genitival expressions (e.g. *hiesig* corresponding to *hier* 'here', *derzeitig* corresponding to *der Zeit* 'of the time'), which are quite absent in English: there is no adjective for *here*, and a phrase such as *die derzeitigen Verhältnisse* has to be translated as *the circumstances at that time*. Some lexical gaps are the result of productive derivational morphology in one language which is not parallelled in another. A good example is the Dutch phrase *het Turks kennen* 'to know Turkish', from which the corresponding nominal *kenner van het Turks* can be formed. But in English the equivalent nominalization **knower of Turkish* is not possible, so the Dutch phrase must be translated as *someone who knows Turkish*. The transfer problems caused by such lexical gaps relate not so much to translation from the single lexical item to a periphrastic expression, but translation in the other direction. In the former, it may well be possible to decide on a particular 'standard' rendition; but in the latter, the variety of locutions may be so resistant to precise specification that the single lexical form is not selected when it would be appropriate.

6.2 Structural differences

Many relatively trivial syntactic differences between languages are well known, e.g. in French most adjectives follow nouns but in English adjectives normally precede the nouns they qualify. Other familiar differences are those between languages such as Japanese and Latin where the main (finite) verb of a sentence comes last and languages such as English and German where the finite verb comes after the first noun (phrase). However, there are many instances where it is less easy to equate a structure in one language with a structure in another in quite such a simple way.

A good example is the passive. In English, this is defined as any construction formed with the auxiliary verb *be* and the past participle. Even when translating into a language which has a similar construction, it is not always appropriate — or even possible — to use that construction. In (5a) the corresponding construction is acceptable; but in (5b,c), alternative constructions are preferable.

(5a) *Le bâtiment fut construit en 1923.*
'The building was constructed in 1923.'

(5b) *Ces livres se lisent facilement.*
Lit. These books read themselves easily
'These books are easily read'

(5c) *Ici on parle anglais.*
Lit. One speaks English here
'English is spoken here.'

Because all these constructions can (sometimes) be used to translate an English passive, it is sometimes said — misleadingly — that there are three forms of the passive in French. However, the reflexive and impersonal constructions also correspond to other constructions in English. Further complications are introduced by the fact that in English it is possible to passivize indirect objects and prepositional objects, which is impossible in French (6).

(6a) Mary was given a book.
**Mary fut donné un livre.*

(6b) This bed has been slept in.
**Ce lit a été dormi dans.*

Some languages, such as German and Japanese, allow the passivization of intransitive verbs (7), which is not possible in English. It should also be noted that the interpretation of this construction is different in the two languages.

(7a) *Es wurde getanzt.*
Lit. It was danced.
'There was dancing.'

(7b) *Toshio-ga tsuma-ni nigerareta.*
Lit. Toshio was run away by his wife.
'Toshio's wife ran away on him.'

Another distinction is found in German, which has two forms of the passive expressing either a process or a completed state, using the auxiliaries *werden* or *sein*, respectively (8).

(8a) *Das Fenster wurde gebrochen.*
'The window was (i.e. got) broken.'

(8b) *Das Fenster war gebrochen.*
'The window was (i.e. had been) broken.'

These examples illustrate the diversity of both forms and functions of the 'passive'. Formal differences include verb morphology, different word orders, use of auxiliaries, alternative case markings on nouns, or combinations of all of these. The functional differences include thematization, depersonalization, and so on. Translation of such constructions is not a simple matter of deciding on structural equivalences.

Other examples are linked closely to particular lexical items or semantic fields. Consider the English and German translationally equivalent sentences in (9):

(9a) I like swimming.

(9b) *Ich schwimme gern.*

The main verb in English, *like*, expresses the 'pleasure' concept, with the activity itself, *swimming* as a dependent participial complement. In German, the main verb expresses the activity *schwimmen* ('swim'), with the notion of pleasure conveyed by the dependent adverb *gern* ('gladly'). Translation from English into German or German into English involves therefore a change of structure, either direct or indirect (i.e. via an interlingual representation, see below). A similar phenomenon is found in Spanish where the verb *soler* which takes an infinitival complement corresponds to the English adverb *usually* (10). A comparable translation problem can be found for almost any language pair.

(10a) *Juan suele ir al cine los sábados.*

(10b) Juan usually goes to the cinema on Saturdays.

Differences of structure within the same semantic field may be illustrated by the English and French equivalents for certain expressions of movement (11)–(13).

(11a) He walked across the road.

(11b) *Il traversa la rue à pied.*

(12a) She drove into town.

(12b) *Elle entra dans la ville en voiture.*

(13a) They flew from Gatwick.

(13b) *Ils partirent par avion de Gatwick.*

In each of the English sentences it is the mode of transport that is expressed by the verb (*walk, drive, fly*) and the direction is expressed by prepositions (*across, into, from*). In the French, it is the direction which is expressed by the verb (*traverser* 'cross', *entrer* 'enter', *partir* 'leave'), and the mode of transport is expressed by adverbial adjuncts (*à pied* 'on foot', *en voiture* 'by car', *par avion* 'by plane').

Another more complex illustration is found in sentences expressing 'naming' (14).

(14a) His name is Julian.

(14b) *Er heißt Julian.*

The English sentence uses the equative (copula) *is* with two arguments: the word *name* modified by the possessive adjective *his*, and the name itself *Julian*. In German the predicate *heißt* ('is named') has two arguments, the person named *er* ('he') and the name itself *Julian*. Further variations are found in other languages: for example in Russian a three-argument construction is used (14c) and in French there is a reflexive construction (14d).

(14c) *Jego zovut Julian.*
Lit. 'They call him Julian'

(14d) *Il s'appelle Julian.*
Lit. 'He calls himself Julian'

These are just a few of many grammatical phenomena we could have chosen to illustrate problems even — apart from (7b) — within a single language family group. Others include the interpretation of verb tenses (where different languages have differing ways of expressing temporal location, including systems of morphological tenses, uses of temporal adverbials, various auxiliaries); modality (expressing necessity, obligation, ability, intention, desire, and so on); quantifier scope (already alluded to in section 5.5); determiner systems (*the*, *this*, *that*, *a*, *some* etc.); and even systems of singular and plural.

One area of particular interest in MT is the translation of **idioms**. The term 'idiom' is rather over-used, and tends to cover everything from fixed phrases like *raining cats and dogs* or *kick the bucket* to slightly metaphorical uses of words such as *pay attention*. It should be said that texts which are rich in colourful idioms and flowery language are probably not ideal material for translation by machine. However, many MT researchers have made considerable efforts to tackle the problem. From the point of view of translation, 'idiom' can be defined functionally. A phrase that can be translated **compositionally**, even if clearly idiomatic in meaning, need not be treated as such (e.g. *to ask for the moon*, French *demander la lune*). On the other hand, any phrase for which the translation is not an obvious combination of the translation of its components must be treated in a special manner (as an 'idiom'). So phrases like *commit suicide* require special attention. Both the verb *commit* and the noun *suicide* have their equivalents in French (*commettre*, *suicide*), and indeed the direct object of *commit* is usually translated as the direct object of *commettre* as in *commit a crime* → *commettre un crime*. But if the direct object happens to be *suicide*, then the whole phrase must be translated differently, by the reflexive verb *se suicider*.

Idioms often occur in constructions which change their structure. Two examples will illustrate this. In French, a phrase like *donner un coup de poing* (literally 'give a blow with the fist') can be treated as an idiom since it translates into English as *punch*. However, in French the construction is a quite regular one — verb plus direct object — which can be augmented by insertion of an indirect object or an adverbial (15a), by modification of the direct object (15b) or by coordination with another similar construction (15c). All these are special problems for an MT system.

(15a) *Elle donna de temps en temps à son frère un coup de poing.*
Lit. She gave from time to time to her brother a blow with the fist.
'She punched her brother from time to time'

(15b) *Elle lui donna quelques coups de poing violents.*
Lit. She gave him several violent blows with the fist.
'She punched him violently a few times'

(15c) *Elle lui donna des coups de poing et de pied.*
Lit. She gave him blows with the fist and foot.
'She punched and kicked him'

The second example comes from Japanese, where the English verb *rain* corresponds to an expression meaning 'rain falls'. If the words occur in a straightforward finite clause, then translation by the English verb is usually correct (16a). But as a regular subject-plus-verb construction in Japanese, it can also be nominalized (16b) or made into a relative clause (16c). In these cases the translation requires special treatment.

(16a) *Kinō ame ga futa.*
Lit. Yesterday rain fell
'It was raining yesterday'

(16b) *Ame no furi kata wa odorokashita.*
Lit. The rain's way of falling surprised us
*'Its way of raining surprised us'
'It surprised us how (hard) it was raining'

(16c) *Kinō furi hajimeta ame ga mada futte iru.*
Lit. The rain which started to fall yesterday is still falling.
*'It which rained yesterday is still raining'
'The rain which started yesterday is continuing'

6.3 Levels of transfer representation

In describing the strategies for MT systems (Chapter 4), it was noted that the basic differences lie in the relative sizes of the three components: analysis, transfer and generation. The direct method stands at one extreme, the interlingua method at the other, with transfer-based systems between them. The well-known 'pyramid' diagram in Figure 6.1. is often used to illustrate this point.

The diagram shows source language analysis up the left-hand side, and target-language generation down the right. The apex of the pyramid represents the theoretical interlingual representation achieved by monolingual analysis and suitable for direct use by generation. However, the path to that interlingua is long, and, as the diagram is supposed to show, by cutting off the monolingual analysis at some point and entering into a bilingual transfer phase, one can avoid the difficulties of a full analysis (as described in Chapter 5). The diagram is also intended to suggest that the more the text is analysed, the simpler transfer will be (as depicted by the length of the line cutting across the pyramid). The extreme case is at the very bottom, where there is minimal monolingual analysis, and nearly all the work is done in transfer, as was the case with the early direct method systems.

To illustrate the various levels of transfer in these different types of systems, we describe the kinds of analysis that could be required, working 'upwards' from the base of the pyramid to the apex. In most cases, representations will be tree

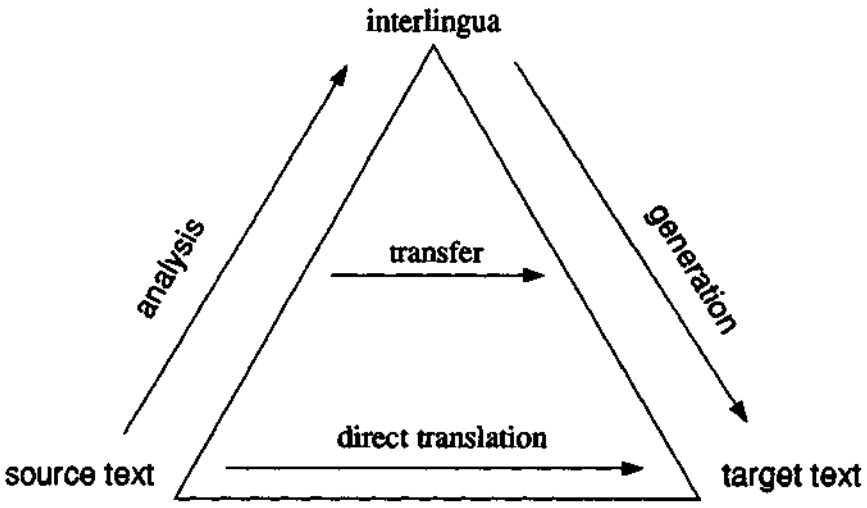


Figure 6.1 Transfer and interlingua 'pyramid' diagram

structures, and typically carry rather more information than is often shown when linguists discuss parse-trees or deep-structure representations (cf. Chapter 2). Our illustrative sentence will be sentence (17) and the target language French.

(17) Any government is dependent on its supporters.

In what follows, we take a deliberately eclectic and varied approach to the linguistic analysis. The reader should not be too concerned about the details of the representations, except to see what new information is introduced at each 'deeper' stage. It should be stressed that no specific method of analysis or representation is being especially promoted by us in this exercise.

6.4 Morphological transfer

The shallowest of analyses might be a word-by-word morphological analysis, resulting in a structure like (18).

(18)	any	government	be	dependent	on	it	supporter
	cat=det	cat=n num=sg	cat=v num=sg pers=3 tns=pres	cat=adj	cat=prep	cat=pospron num=sg pers=3 sex=neut	cat=n num=pl

The analysis identifies grammatical categories, the syntactic number (singular or plural) of nouns, and the tense of verbs, but there is no identification of relations between words (e.g. determiners and nouns) or of groupings (noun phrases). In a crude word-for-word system transfer would consist simply in the substitution of source language words (with category features) by target words (with corresponding features), as in (19a), with perhaps the resulting target text (19b).

(19a)	quelconque cat=det	gouvernement cat=n num=sg	être cat=v num=sg pers=3 tns=pres	dépendant cat=adj	sur cat=prep	son cat=pospron num=sg pers=3 sex=neut	défenseur cat=n num=pl

(19b) **Quelconque gouvernement est dépendant sur son défenseurs.*

The resulting translation would be clearly inadequate, even for a sentence such as this where little reordering is needed. It is relatively easy to think of English sentences — (20a) for example — for which this level of analysis results in almost incomprehensible and seriously ungrammatical output (20b).

(20a) Any responsible and well administered organisation must look after its female employees.

(20b) **Quelconque responsable et bien administré organisation doit regarder après son féminin employés.*

In practice, most of the direct translation systems do include minimal identification of local context so that the English sequence adj+noun would be inverted for French. However, because some French adjectives precede nouns (e.g. *bon, jeune*), exceptions would have to be indicated in the dictionary entries for these items.

These kinds of minimal structural changes are sometimes incorporated in a separate phase of direct systems which may operate after lexical substitution (cf. Figure 4.1): the arrangement is found particularly in direct systems with some modularity (e.g. Systran, Chapter 10). For this reason, 'local reordering' is often regarded as the beginning of 'generation' in these systems (see section 7.1 below.)

Example (20) illustrates another problem for word-for-word translation. The phrasal verb *look after* should be treated as a unit. Clearly it could be entered as a compound in the dictionary (with the French equivalent *soigner*). However, other phrasal verbs can be split (21).

(21a) He looked up the number in the directory.

(21b) He looked the number up in the directory.

With no identification of verb phrase structure, such compounds cause difficulties. How can *up* be related to *look*? And in (22), how can it be decided when *in* is part of the phrasal verb *fill in* and when not?

(22a) He filled it in.

(22b) He filled it in haste.

(22c) He filled it in yesterday.

(22d) He filled it in the morning.

The direct approach has particular problems with homographs; the usual method of resolving homograph ambiguities is to look at adjacent words for clues. For example, in the case of *empty*, which may be an adjective or a verb, it is the grammatical categories of adjacent words which are looked at: if *empty* comes between a determiner and a noun then it is presumed to be an adjective. But it may also come between another adjective and a noun, or at the end of a sentence. Listing all the possible category adjacencies for every adjective/noun

and noun/verb homograph is clearly unsatisfactory, but this in essence is what direct systems have to do.

In the case of a word such as *enter* which may have a number of different translations (*entrer, s'enrôler, inscrire*) the program has to search for specific words (e.g. *room, service, ledger*). But there are obvious limits (computational, for example) on how much context can be examined, even if all the relevant words can be listed. Furthermore, there can no assurance against misleading contexts (23)

(23) The cleaner entered the service room.

However, even this approach — cumbersome as it is — is useless in cases such as the choice between *savoir* and *connaître* to translate *know*; only by examining the structural environment can a decision be reached (section 6.1 above). But without a structural analysis nothing can be done. The only solution, as many earlier direct systems found, is to give alternative translations and leave the choice to post-editors (section 8.3.2).

Obviously, direct systems hope that few structural and lexical choices have to be made. The more 'free rides' (section 5.3.2.4) that can be found, the simpler the dictionary entries can be. It is a reasonable assumption that fewer structural changes are required between cognate languages, such as French and Italian, than between unrelated languages, such as English and Japanese. Nevertheless, the paucity of structural analysis can be overcome only by the *ad hoc* inclusion of multiple-word entries, resulting in a huge bilingual dictionary containing structural transfer rules hidden in lexical transfer rules.

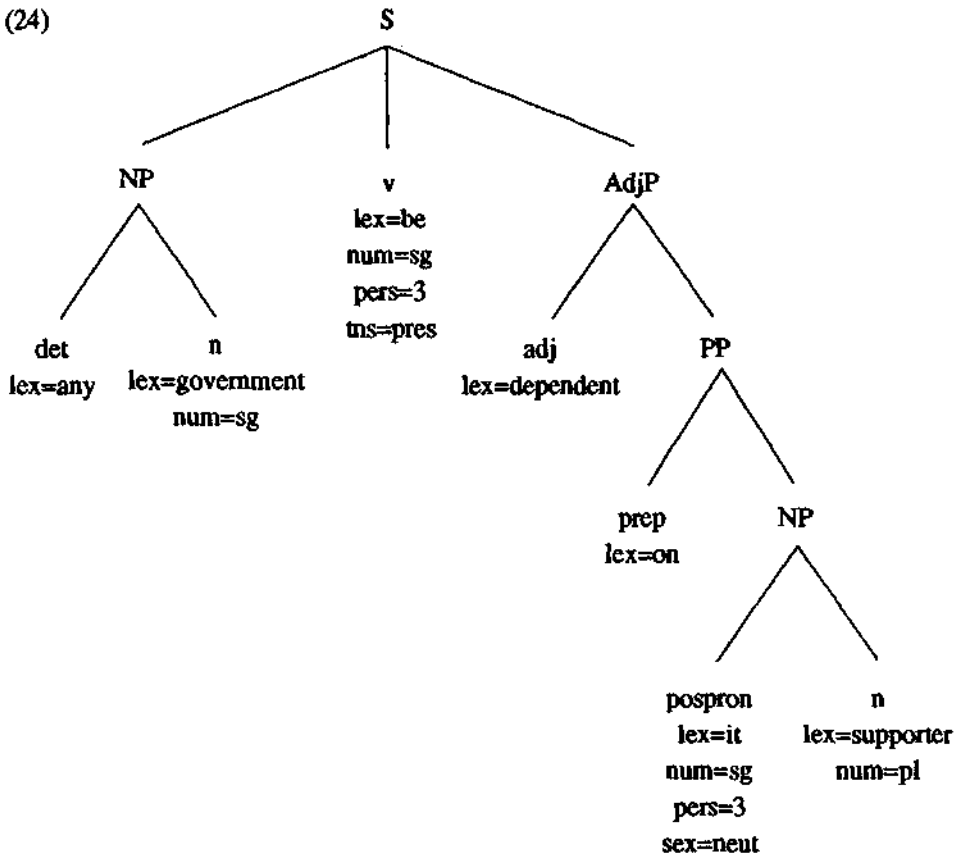
One answer to the problem of excessive lexical and structural detail in the dictionary is to conduct statistical analyses of texts and thus to derive probability measures for converting source language words or expressions into target words or expressions. No grammatical, semantic or lexical information need be involved, only information on the probabilities of source words corresponding to target words. The statistical approach was favoured by a number of the earliest MT systems (cf. section 1.3), and has now been revived by a group at the IBM Laboratories at Yorktown Heights, NY. Their tests of the approach on a large bilingual corpus of Canadian parliamentary proceedings have demonstrated that the use of probabilistic correspondences can produce context-sensitive translations with some success, with little or no use of grammatical information; the results have naturally aroused considerable interest (section 18.3).

6.5 Transfer-based systems

The next level of analysis after (18) is given by a syntactic parse resulting in a surface-structure representation. For sentence (17) this might result in a parse tree like (24) (next page).

Although there is now some structural information, we still do not have any information about functional relationships between elements. Using this structure as a basis for transfer into French, we might get internally coherent constituents, but we would not know what kind of structure the sentence as a whole should have, since there is no analysis of the syntactic or semantic functions of the elements.

(24)



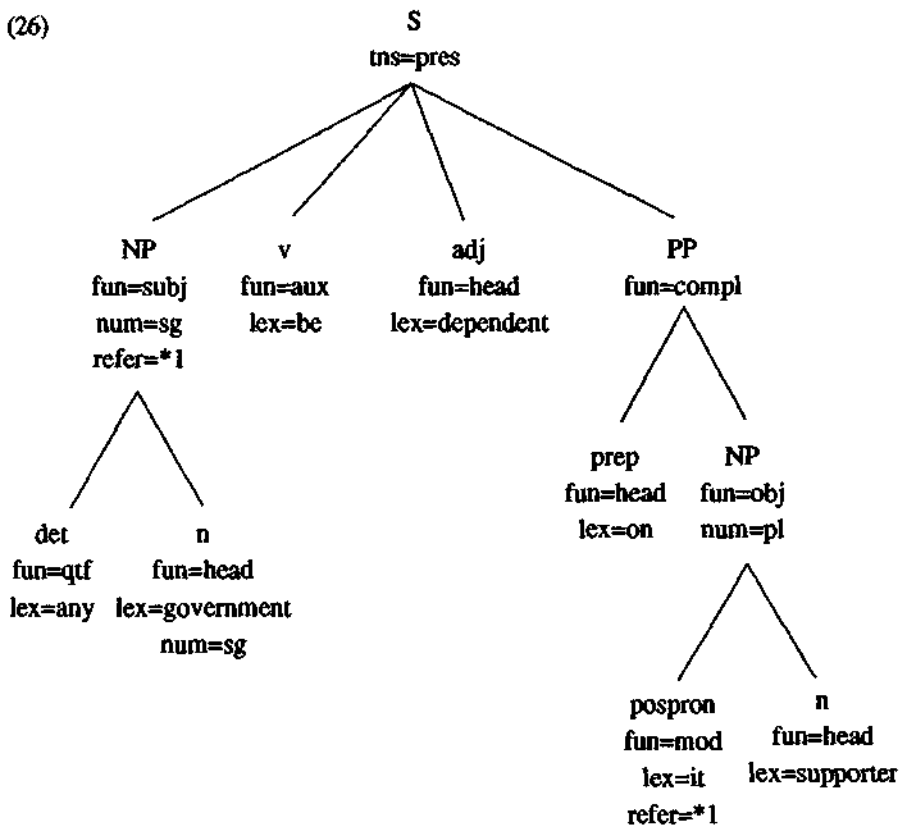
Nevertheless, some early transfer-based systems assumed that successful structural transformations between languages could be achieved at this 'depth' of analysis. For our example sentence (17) and for translation into French it might be sufficient. But in a relatively simple three-argument sentence such as (25a) it is essential to distinguish direct object and indirect object. This is even more obvious in a case like (25b) since structural changes are necessary.

(25a) The professor gave the student an assignment.

(25b) The student gave the teacher no offence.
L'étudiant n'a pas offensé le professeur.

It is now generally accepted that transfer must be based on a 'deeper' analysis, one which includes at least some indication of functional relations. For (17) the result might be as in (26), in which the syntactic functions of the constituents have been recognised, especially that the finite verb *be* is an auxiliary here, and the main predicate is the adjective. Other functions include subj[ect], compl[ement], qtf

(26)



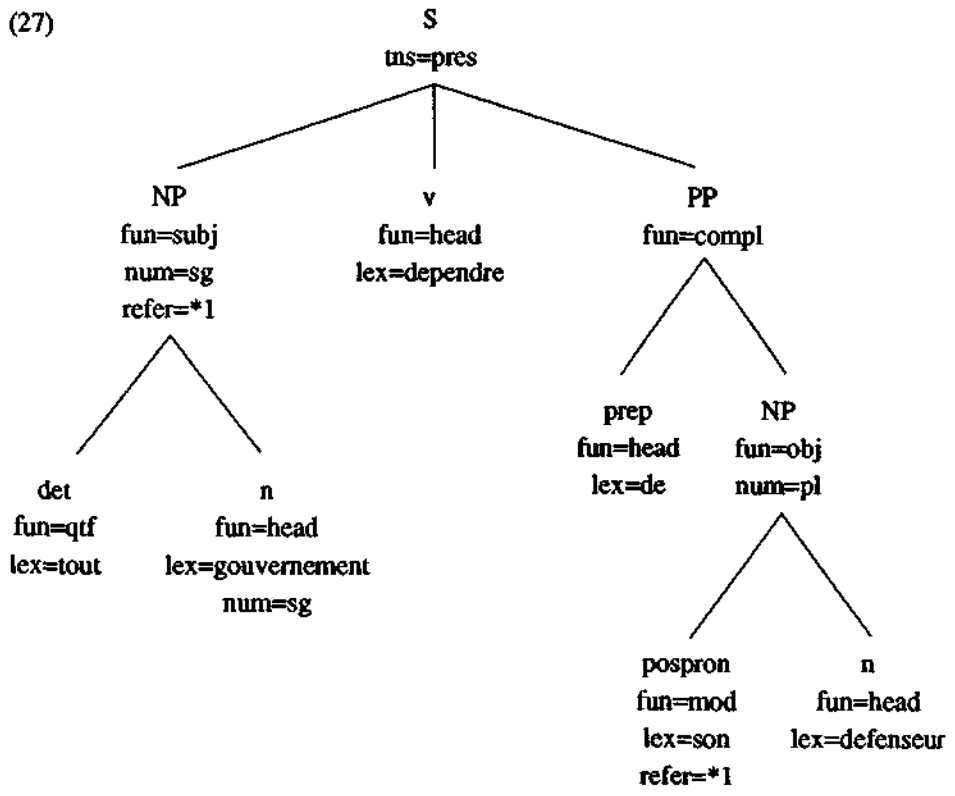
(quantifier), mod[ifier], and, for each constituent, head. Grammatical information such as tense and number has also been passed up to the highest appropriate nodes.

Notice that by matching up values for 'sex' and 'num', it is possible to recognise the anaphoric relation between the subject NP and the possessive pronoun, and to assign to both the refer(ence) index '*1'.

What is now involved in the transfer of this structure into French? There are two basic aspects: lexical transfer and structural transfer. Both these processes can cause problems (sections 6.5.1 and 6.5.2 below), but here for simplicity we may assume that lexical transfer consists of word-for-word replacement. The translation of the quantifier *any* might well be difficult — it has to be *tout* ('all') (see section 7.2.2). We know that the referent of the possessive pronoun (*its*) is the subject NP (*government*). For French the important question is whether it is singular or plural, since French possessive pronouns do not reflect the sex of the antecedent (i.e. 'his' and 'her' are the same), though they do agree in gender with the noun they modify. In this case, then, it so happens that it would not matter if the analysis of the anaphora was incorrect or incomplete: for translation into French, this is a 'free ride' (see 5.3.2.4 above). To transfer the preposition, we have to check which constituent the PP is attached to. Because the PP is a complement, its translation is determined by the head of the constituent (cf. *depend on*, *look for*,

finish with etc.); by contrast, the translation of an adverbial PP often depends more on the object of the preposition itself (cf. *on the hill, in the city, at the bank* if the PP is a locative adverbial). The main structural change involves the auxiliary-plus-adjective, which is to be a straightforward verbal structure in French, i.e. *be dependent* is translated as *dépendre*. The result of these transfer operations is an input representation for French generation such as (27).

(27)



This particular illustration of transfer is relatively straightforward, but one could well imagine having to change the PP complement into a direct object, or having to make other basic structural changes (cf. some of the examples given in 6.2 above). This suggests perhaps that this level of representation is still a little too close to the English surface structure, and transfer might be somewhat easier if we could make our representation a little deeper, as we shall illustrate below.

Most transfer-based systems perform transfer on intermediate (or 'interface') representations at a level something like that illustrated in (26) and (27), as we shall see in later chapters dealing with actual systems. The main point to notice is that the representations remain language-specific: they typically have source and target language words in them, and they reflect the structure, whether superficially or in more depth, of the respective source and target languages. By contrast, interlingua-based systems attempt to provide representations which are language-independent both in lexicon and in structure.

Before discussing interlinguas, we must look more closely at problems of lexical and structural transfer in transfer-based systems.

6.5.1 Lexical transfer

Lexical transfer consists of the replacement of a source lexical item by a target lexical item. Obviously if there is only one target language equivalent (*north*, French *nord*, German *Nord*; *library*, French *bibliothèque*), then there is no problem, except where there are attendant differences of syntactic structure (see below). But only in the area of technical translation may such one-to-one lexical correspondences (of terms) be expected to be common: *screen – écran*; *hæmoglobin – hémoglobine*; *data processing – Datenverarbeitung*). Equally unproblematic for lexical transfer are many-to-one translations, e.g. German *Mauer* and *Wand*, both translate as English *wall*. The problems arise with the numerous one-to-many translations of lexical items (cf. the examples in section 6.1 above). In some circumstances it may be possible for an MT system to ignore certain translations, because they are rarely used or outdated (e.g. *computer* could be *Rechner* in German, but currently the more usual term is *Computer*), or because one of the possibilities occurs only in a particular sublanguage (section 8.4).

Where choices have to be made it is normal for lexical transfer to require inspection of surrounding context. This is the case with translating *know* (section 6.1) where translation into French or German can be determined from the identification of the complement as a clause or a noun phrase. Another example involves translating *eat* into German: *essen* is chosen if the subject is human, but *fressen* otherwise; in other words, the 'subject' or Agent must be identified and its semantic features checked for compatibility with the selection restrictions of the German verbs. More difficult are translations where the context is less easily defined in terms of adjacent vocabulary: a *library* is in German *Bibliothek* if it is part of an academic or research institution, but *Bücherei* if it is open to the general public. Many examples were given above of cases where for lexical transfer it is necessary to make a distinction that may be obvious to a human but which is difficult to express in terms of local contextually available information.

6.5.2 Structural transfer

Structural transfer is necessary when the structure inherited from the source language is inappropriate for the target language. In theory, the deeper the analysis goes, the less likely this problem is to occur, since the deepening analysis aims at neutralising the distinctions between languages.

Some problems of structural transfer are easier to solve than others. Consider a sentence like (28a) and its translation into French or German.

(28a) Jones likes the film.

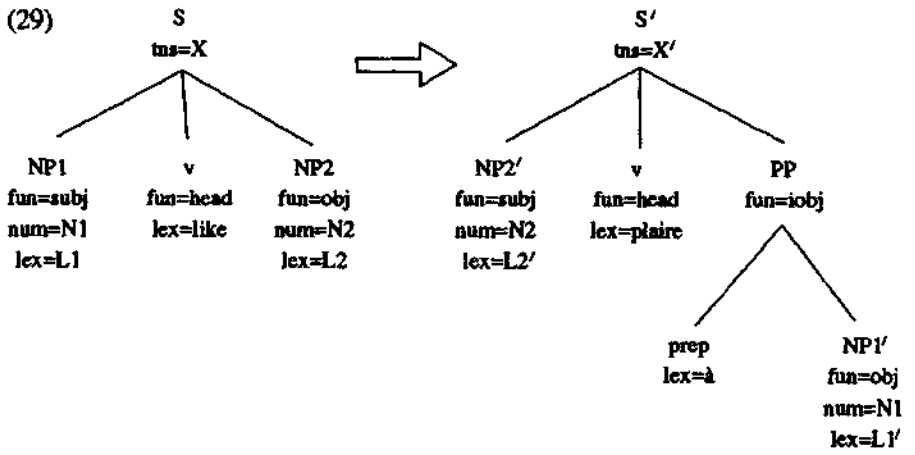
(28b) *Le film plaît à Jones.*

(28c) *Der Film gefällt dem Jones.*

While *like* can be said to translate as *plaire* or *gefallen*, the corresponding structures do not match, since the subject of *like* has to be mapped onto the

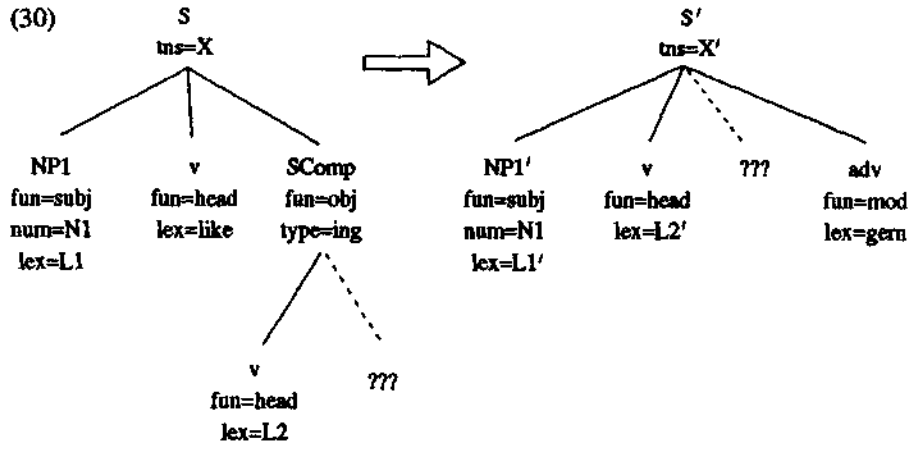
indirect object of *plaire* or the dative object of *gefallen*, while the object of *like* becomes the subject in both cases. This can be treated as a structural 'transfer rule' formulated in a quite straightforward way, perhaps something like (29) for English to French (with other transfer rules implied by the correspondence of similarly labelled elements, e.g. NP2 → NP2′).

(29)



Rather more complex would be the 'transfer rules' for the examples above involving *like* and *gern* (9), verbs of motion in English and French (11-13), and naming verbs (14). All involve substantial structural changes. The partial example rule (30) gives a flavour of this complex structural transfer for *like* and *gern*.

(30)



Notice that all the arguments of the embedded verb (indicated by the dotted line and '???') — which may also include further adverbials as in (31a) — have to be somehow transferred up into the top level structure.

(31a) John likes swimming with his friends in the summer.
John schwimmt im Sommer mit seinen Freunden gern.

It should also be noticed that rule (30) has to specify that the SComp contains a present participle (type=ing), since in other constructions *like* can take an infinitival SComp (as in 31b,c) where the transfer rule would again be different.

(31b) John likes his father to play cricket with him.

(31c) John likes to play cricket with his father.

Another example of the necessity for structural transfer comes from Japanese. The translation of sentences like those in (32) into Japanese involves a rule known as the 'animate subject constraint'.

(32a) The wind opened the door.

(32b) The earthquake destroyed the buildings.

(32c) A short walk will bring you to the station.

(32d) Our limited budget will not allow us to start a new project.

This constraint requires the subject of a verb which expresses an action in Japanese to be animate. This does not mean that all action verbs must have an animate subject, since in Japanese it is possible to omit the subject altogether, but formulations like those in (32) are not possible, because in each case the subject is not animate. Instead, the structure of the sentence must be transformed so that the implied subject is expressed, or omitted altogether, and the element which was the subject in the English must be expressed as an adverbial of the appropriate type. So the Japanese translations of the sentences in (32) would be structurally more like the English equivalents in (33).

(33a) By means of the wind the door became open.

(33b) The buildings were destroyed because of the earthquake.

(33c) By walking a short way, you will arrive at the station.

(33d) Owing to our limited budget, we cannot start a new project.

As a final example of structural transfer, we look at nominalizations in English, where it is generally possible, and even stylistically preferred, to construct sentences consisting of complex nominalizations connected by relatively vague verbs like *be*, *do*, *have*, rather than use relative clauses involving verbs. This is particularly the case in scientific writing: consider the examples in (34).

(34a) The possibility of rectification of the fault by the insertion of a wedge is discussed.

(34b) There has not been a substantial change in Tokyo residents' consumption of reservoir water despite the announcement by the Government of a period of rationing.

(34c) The extraction of moisture from the air by the dehumidifier made the room more comfortable.

In some languages, however, such constructions, even if they are grammatically possible, are considered very clumsy, difficult to understand, or anglicizations. Japanese and Czech are two such languages, and structure-preserving translations of the above sentences (i.e. translations which reflect the preference for nominalizations), while grammatically correct, would be regarded as stylistically inferior. Structural transfer is required to 'denominalize' the constructions, that is to say locate the underlying verb, identify the functional relationships of the

components, and generate from there. The preferred structures would be more like the English equivalents in (35).

(35a) We discuss whether it is possible to rectify the fault by inserting a wedge

(35b) The amount of reservoir water that Tokyo residents consume has not substantially changed, even though the Government has announced a period during which it will be rationed.

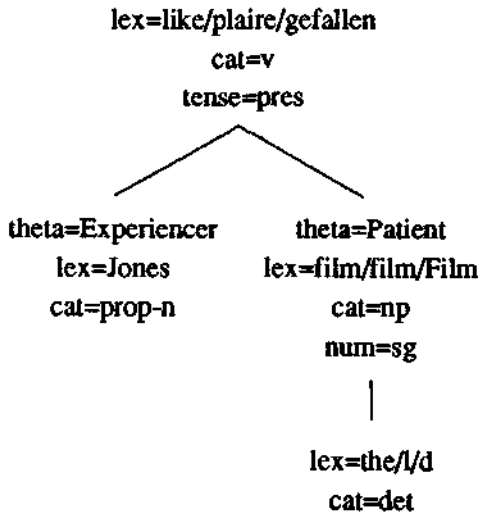
(35c) When they used the dehumidifier to extract the moisture from the air, it became more comfortable in the room.

6.6 Transfer with a structural interlingua

The problems of structural transfer underline the problems with representations which do not neutralize sufficiently the idiosyncracies of even the deep structures of the respective languages. As mentioned in section 2.9.2, it was once thought that the 'deep' structures of transformational-generative grammar might serve as interlingual representations, since they were held to be language-independent bases for semantic interpretations. A more common basis for an interlingual representation is the use of some of the ideas found in Case grammar.

Case roles (also known as 'deep cases', 'semantic roles', 'thematic roles' or 'theta roles') are semantic relations such as Agent, Patient, Experiencer, Instrument, etc., as discussed in section 2.9.6. Some of the less complex problems of structural transfer disappear when this sort of representation is adopted. Examples like (28) above no longer involve complex transfer, since the Case-based representation for both languages can be identical, except for the lexical items (36). Note the usual preference with a Case grammar analysis for a dependency representation rather than a phrase structure tree.

(36)

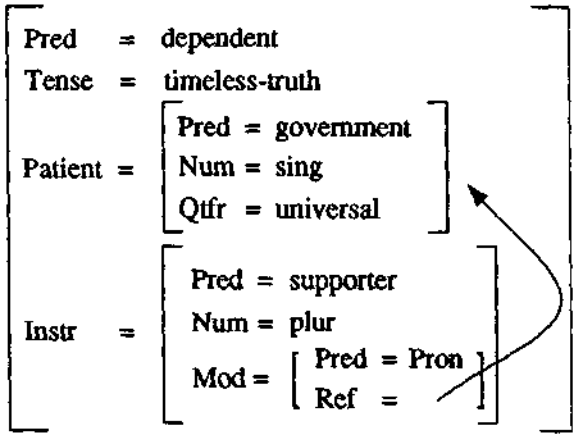


As an interlingual representation, this would be suitable as input to French or German generation, which will derive the appropriate surface structure from the representation, i.e. selecting the Patient as the surface subject, and the Experiencer as the indirect object (and making the necessary selections for tense and gender agreement).

Case structure representations are frequently supposed to reflect universals of syntax which could be regarded as interlingual. Use of Case roles is widespread in bilingual transfer-based MT systems, particularly when one of the languages is Japanese since Case grammar seems to be more appropriate as an analysis model for this language than the kind of phrase structure model often applied to European languages (cf. the examples in (32) where a Case-based analysis would identify the underlying roles of the surface subjects in the English sentences). Unfortunately, despite intensive study by linguists, there is no widely agreed set of Case relations, so researchers in MT systems adopting this approach are generally obliged to devise their own set of Case roles.

Returning to our example sentence (17), it will be recalled that the deepest level of representation so far proposed (26) still left us with some structural transfer regarding the change from the auxiliary-plus-adjective structure to the verb-plus-prepositional object structure of the French. This structural transfer could be avoided by the use of a Case-based representation as in (37). It is convenient here to adopt a feature-structure representation as seen in section 2.8.3.

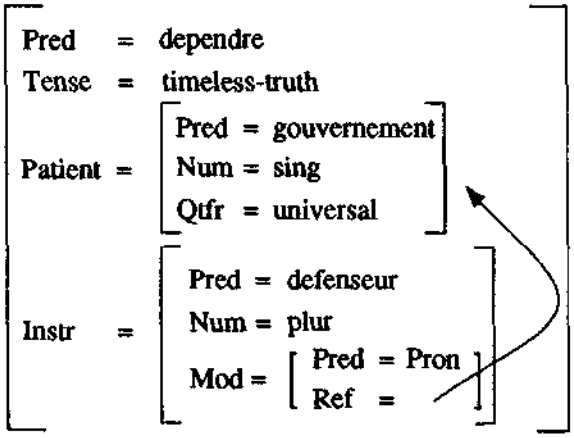
(37)



Some of the relationships represented in (37) have already been recognised in (26) and (27), e.g. the co-reference of the possessive pronoun (the Mod of *supporter*) and the Patient of the main predicate — we show here an alternative way of representing co-reference (an explicit 'pointer' instead of the co-indexing 'refer=*1' in (26) and (27)) — either method can be used in these representations. In other aspects the analysis is deeper, reflecting the recognition of *be dependent* as a single predicate, the description of *any* as a 'universal' quantifier, the suppression or replacement of syntax-oriented information such as category, lexical content of

minor categories (determiner, possessive pronoun), and grammatical tense, all of which can all be recovered from the other more semantic information contained in the representation. Transfer to the equivalent French (38) is now little more than lexical replacement of *dependent* by *dépendre*, *government* by *gouvernement*, and *supporter* by *défenseur*. All other lexical items in the target text would be derived during generation (see next chapter).

(38)



With structures such as (37)–(38) we are close to an interlingua-based representation, but only in respect to structural features; lexical transfer must still be done bilingually. As it happens, the conversion of (37) to (38) does not involve any 'translational ambiguity', but if instead of *government* we had *party* then a choice between *parti* (political party), *groupe* (party of travellers etc.), *partie* (legal term, as in *third party*) and *fête* ('celebration') would have to be made.

6.7 Interlingua-based systems

In interlingua MT systems, the result of source language analysis is a language-independent representation of the text which is the basis for the generation of the target language text. The advantages for multilingual systems have been mentioned already (section 4.2). The disadvantages arise from the fact that analysis and generation have to be strictly separated; it is not desirable to orient analysis towards a particular target language and it is not possible during generation to look back at the original source language text. The interlingua representation has to include all the information that might conceivably be required during the generation of any target language text (or rather more precisely: any target language included in the system now or intended to be added in the future). In effect this high degree of language-independence and neutrality means that interlinguas must strive towards universality in lexicon and structure: one might almost say, towards representing the 'meaning' of the text.

6.7.1 Structural representation in interlingua systems

One candidate for interlingua representation is the use of the formulae of propositional logic, and there has been some experimentation in this area. Taking once again our example sentence (17), a representation based on logical formulae might be something like (39).

(39) `all(X), government(X), indefinite(Y), plural(Y), support(Y,X,T), depend-on(X,Y,T), timeless(T)`

Research has shown that it is generally not too difficult to arrive at this sort of representation given an input sentence such as (17). More difficult however is the generation of text from such a representation. For example, the sentences in (40) are only a selection of the possible texts that could be generated (in English) from (39), and it seems that the total absence of information relating to syntactic structure (e.g. indication of topic and comment, or choice of main predicate) makes such a representation inappropriate for translation (though well suited to multilingual **paraphrasing**, if that is an acceptable alternative to translation). In fact, (some of) the paraphrases in (40) could be said to distort the original message. Furthermore, no account has been taken of alternative surface forms for predicates such as 'depend-on(X,Y,T)' or 'support(Y,X,T)'. (It must be stressed that the use of something that resembles an English lexical item as the name of a logical relation should not be taken as an indication of a one-to-one relationship between the relation and the corresponding word.)

(40a) Every government has supporters which it depends on.

(40b) People who support all governments are depended on by them.

(40c) All governments depend on people who support them.

Most interlingua-based systems use representations which are essentially structured like those seen in the previous section for transfer-based systems: the difference is in the abstractness of the representation of structure, and in the treatment of lexical items.

There is general agreement among proponents of the interlingua approach that representations should be hierarchies of substructures showing clear inter-relationships. As already mentioned, in the earlier days of MT research on interlinguas, the Chomskyan theory of deep structures was thought to be attractive, but it is now agreed they are not sufficiently abstract, being too oriented towards the surface features of individual languages. Case grammar was also considered, but as already remarked, there is no agreement yet on a 'universal' set of case roles.

The implications of neutral structural representations can be illustrated by considering differences of word order between languages, and their significance. In English, word order is the primary means of distinguishing grammatical functions like subject and object. In (41), word order alone determines which noun phrase is the 'chaser' (the subject), and which is the 'chased' (the object).

(41a) The man chased the shark.

(41b) The shark chased the man.

Languages such as German and Japanese indicate grammatical functions with overt case-marking, so that word-order differences have other functions: in (42a)

and (42b) the man is the topic of discourse (cf. section 2.7), while in (42c) it is the shark. But the underlying event described in (42b) and (42c) remains the same.

(42a) *Der Mann jagte den Haifisch.*
THE-NOM MAN CHASED THE-ACC SHARK
'The man chased the shark'

(42b) *Den Mann jagte der Haifisch.*
THE-ACC MAN CHASED THE-NOM SHARK
'It was the man that the shark chased'

(42c) *Der Haifisch jagte den Mann.*
THE-NOM SHARK CHASED THE-ACC MAN
'The shark chased the man'

Similarly in Japanese (43):

(43a) *Same ga hito wo oikaketa.*
SHARK-NOM MAN-ACC CHASED
'The shark chased the man'

(43b) *Hito wo same ga oikaketa.*
MAN-ACC SHARK-NOM CHASED
'It was the man that the shark chased'

In Russian, another highly inflected case-marking language, word order often indicates differences between definiteness and indefiniteness, i.e. between what is 'given' or 'presupposed' and what is mentioned for the first time (cf. section 2.7), as in (44).

(44a) *Ženščina vyšla iz domu.*
WOMAN-NOM CAME OUT HOUSE-GEN
'The woman came out of the house'

(44b) *Iz domu vyšla ženščina.*
'A woman came out of the house'

The implication for an interlingua is that it is not enough to indicate word order on its own: the interlingua must represent the significance in terms of syntactic relations (grammatical function), topic-focus relations (text function), determination, case role or whatever else the interpretation of the word-order dictates.

Passive, as we saw (in 6.2 above), is not a structure which can be equated across languages. In an interlingual representation it is not sufficient to characterise some structure simply as 'passive': rather the representation has to reflect the interpretation of those aspects of the meaning of the sentence as a whole which are indicated by the use of the passive.

Structural differences such as the *like/gern* examples in (9) can be treated in a (relatively) straightforward manner in transfer-based systems by structural transfer rules (section 6.5.2 above). But in interlingua-based systems the representation must be language-neutral. It could be one with an item corresponding to *like* as the head, or it could be one with an item corresponding to *swim*. The first would be a representation more like that of English, the second would be more like German. Either choice would be arbitrary, but neither would be neutral.

Another example was the expression of movement in examples (11)–(13). As we saw in section 6.2 above, in French the verb expresses direction and an adverbial phrase expresses the mode of transport, while in English the verb

expresses the 'motion' and the prepositional phrase expresses the direction. One possible approach for a neutral interlingua representation might be to separate the two elements of 'motion' and 'direction' as in the partial representations shown in (45), where interlingual lexical units are conventionally indicated by angle brackets (for convenience English labels are used):

(45a) He walked across the road
Il traversa la rue à pied

$$\begin{bmatrix} \text{Pred} = \langle\text{MOTION}\rangle \\ \text{Tense} = \text{past} \\ \text{Agent} = \begin{bmatrix} \text{Pred} = \text{Pron} \\ \text{Num} = \text{sing} \\ \text{Pers} = 3 \\ \text{Sex} = \text{male} \end{bmatrix} \\ \text{Instr} = [\ \text{Pred} = \langle\text{FOOT}\rangle\ ] \\ \text{Loc} = \begin{bmatrix} \text{Pred} = \langle\text{CROSS}\rangle \\ \text{Obj} = [\ \text{Pred} = \langle\text{ROAD}\rangle\ ] \end{bmatrix} \end{bmatrix}$$

(45b) She drove into town.
Elle entra dans la ville en voiture.

$$\begin{bmatrix} \text{Pred} = \langle\text{MOTION}\rangle \\ \text{Tense} = \text{past} \\ \text{Agent} = \begin{bmatrix} \text{Pred} = \text{Pron} \\ \text{Num} = \text{sing} \\ \text{Pers} = 3 \\ \text{Sex} = \text{female} \end{bmatrix} \\ \text{Instr} = [\ \text{Pred} = \langle\text{CAR}\rangle\ ] \\ \text{Loc} = \begin{bmatrix} \text{Pred} = \langle\text{ENTER}\rangle \\ \text{Obj} = [\ \text{Pred} = \langle\text{TOWN}\rangle\ ] \end{bmatrix} \end{bmatrix}$$

A frequent option is to adopt the representation which is common to the majority of the languages of the system. For the 'naming' expressions above (14) (repeated here as (46a–d) for convenience) it is not obvious what could be an interlingual structure. However, the preponderance of constructions using a predicate corresponding to 'name' might suggest that underlying structure might

(45c) They flew from Gatwick.
Ils partirent par avion de Gatwick.

```
[ Pred = <MOTION>
  Tense = past
  Agent = [ Pred = Pron
            Num = plur
            Pers = 3 ]
  Instr = [ Pred = <PLANE> ]
  Loc = [ Pred = <LEAVE>
          Obj = [ Pred = "Gatwick" ] ] ]
```

be something like (46e), which happens to correspond most closely in this case to the German (46b).

(46a) His name is Julian.
(46b) *Er heißt Julian.*
(46c) *Jego zovut Julian.*
(46d) *Il s'appelle Julian.*

(46e)

```
[ Pred = <HAS-NAME>
  Tense = present
  Agent = [ Pred = Pron
            Num = sing
            Pers = 3
            Sex = male ]
  Compl = [ Pred = "Julian" ] ]
```

6.7.2 Lexical representation in interlingua systems

The difficulties of devising neutral representations for syntactic structures are matched, perhaps exceeded, by those of establishing neutral representations for lexical units. As with structural differences, the problem for the interlingua approach is again two-fold: deciding on the most appropriate neutral representation, and discovering the procedures for extracting the necessary information from texts. As we have seen, in transfer-based systems there are no problems if for a particular language pair there are one-to-one equivalents; the problems come when there is more than one target word for a single source word. But for an interlingua in a multilingual system there are problems even if only one of the languages involved has two or more potential forms for a single given word in one of the other languages. If an interlingua is to be completely language-neutral, it must represent

not the words of one or other of the languages, but language-independent lexical units, effectively concepts (see also 6.8 below). Any distinction which is (or can be) expressed lexically in the languages of the system must be represented explicitly in the interlingual representation. These distinctions can reflect grammatical, stylistic and most commonly conceptual differences (section 6.1 above).

It is a massive undertaking, even for closely related languages. For example, if Spanish is one of the languages included, the distinctions between the legs of humans, animals, chairs and tables (example (4c) above) has to be made in the interlingua even if when translating between English and German this distinction is irrelevant (*leg*: *Bein*). If the system were to include Japanese the interlingua ought, in theory, to distinguish eight different <WEAR> concepts, which would have to be identified each time the English word *wear* occurred (see examples (4i) above) — even if the target language were German, and the translation in every case would be *tragen*. The alternative would be to have a single <WEAR> concept in the interlingua, and the selection of the appropriate Japanese form would occur during the generation of the Japanese text. However such a unified concept would be unnatural for the Japanese: as unnatural, in fact, as a single concept for both *dove* and *pigeon* (cf. French *colombe*, German *Taube*) would be for English speakers. The situation is further complicated, as already described (in 6.1) by the fact that English and German, but not Japanese, distinguish between the process of 'putting on' and the state of 'wearing' clothes. A true interlingua would have to distinguish sixteen different concepts in this semantic field: <WEAR-HAT>, <PUT-ON-HAT>, <WEAR-COAT>, <PUT-ON-COAT>, <WEAR-GLOVES>, <PUT-ON-GLOVES>, etc.

In practice, nearly all interlingua systems do not do this; instead they rely on contextual information or real world knowledge when it comes to choosing between translation alternatives (see also section 6.8 below). In other words, they revert to a 'lexical transfer' approach with (more or less) interlingual structural representations, as described in 6.6 above.

The exception is to be found in those systems with an interlingua based on an existing language. To be a candidate, such a language has to be unambiguous, consistent and regular in its lexicon, i.e. have no homographs and polysemes, to be regular and unambiguous in syntactic structures, and to be expressive enough to embrace most (ideally all) conceptual differences of languages. No 'natural' languages meet the specification, but artificial languages devised for international communication and actually used like other spoken and written languages are sometimes thought to be possible candidates. One MT project (DLT at Utrecht) has in fact investigated the use as an 'interlingua' of a modified version of Esperanto, which is relatively free of homography and unambiguous in morphology. However, it is doubtful whether a system using a 'natural language' interlingua should strictly be called an interlingua-based system. This is because translation into this type of interlingua from, say, English is itself an MT system; likewise translation from the interlingua into French or another language. These sub-systems themselves can only be 'direct' or transfer-based, even if analysis and generation processes are simpler than when either source or target languages are true natural languages (for more on the DLT interlingua system see Chapter 17).

6.7.3 'Restricted' interlinguas

While a universal interlingua may not be feasible, in particular as far as the lexicon is concerned, we could nevertheless have a representation which is interlingual for exactly the languages we are dealing with. Such an interlingua would neutralise the differences between the languages in the system but take advantage of their accidental similarities. A system could be imagined for a set of Romance languages, for example, covering only French, Spanish, Italian and Portuguese. The interlingua approach might be feasible because these languages have much in common, both lexical correspondences and similarities in grammatical structures. The Eurotra project (Chapter 14) found that it was possible to establish 'euroversals' in certain semantic fields, i.e. interlingual lexical items having common ranges of meaning and reference for the European Community languages concerned.

An interlingua designed in this way for only a small set of European languages becomes even more feasible if we also impose restrictions on the breadth of coverage of such a system, so as to reduce the number of lexical items in the system and the problem of polysemy, or to control the number of syntactic constructions we expect to be able to treat. This 'restricted syntax controlled vocabulary' approach to MT system design is discussed further in Chapter 8.

At this point it is perhaps appropriate to comment on the use of the term 'interlingua' in the MT community. In this discussion, we have taken the term to mean a 'language-neutral intermediate representation'. By 'language-neutral', we have not of course meant that the representation has no linguistic content but only that it is not committed to or does not reflect the surface characteristics of a particular language. As this last section has shown, in practice an 'interlingua' is a representation FOR A GIVEN SET OF LANGUAGES which neutralizes their individual differences. This is of course a theoretically less committed viewpoint than the idea of a 'universal' interlingua, though in practical terms it is probably more tractable; indeed, there is a tendency in the most recent discussions of MT system design towards this conception. Finally, it should be pointed out that the perhaps natural gloss of 'interlingua' as 'intermediate language' can often be misleading. In most cases, an interlingua is a REPRESENTATION, i.e. a data structure in the computational sense of the word, and not a LANGUAGE in the usual sense at all. The exceptions are only the rare use of natural or auxiliary languages (e.g. Esperanto) as interlingual representations. It is therefore misleading for the term 'intermediate language' to be used when referring to the types of interface representations found in transfer-based or indeed in most interlingua-based systems.

6.8 Knowledge-based methods

The discussion so far has concentrated almost exclusively on linguistic approaches to problems of transfer and interlinguas. However, lexical translational differences are also, and inevitably, reflections of differences in cultural backgrounds and differences in conceptual divisions of the same 'reality'. The *rice* example (4h) is a good example of a cultural difference in the culinary sphere; the *wear* example (4i) illustrates a difference in the conceptual partitioning of a universal activity.

It has therefore been argued that translation should be based on non-linguistic conceptual representations, or — in alternative formulations — on representations of meanings derived from the processes of **understanding** of texts. Understanding a text involves relating what is said in a text (its linguistic content) to phenomena (entities, actions, events) outside the text (the non-linguistic 'reality'). Readers interpret texts with reference to what they know about the 'real world' or what worlds they can imagine. On the assumption that meanings and understanding are common and universal to all speakers of human languages, it follows that these 'conceptual' representations are interlingual, that they can serve as intermediate representations in MT systems, and that with the help of suitable 'knowledge bases' texts can be analysed into and generated from such interlingual representations.

We have seen already how reference to **real world knowledge** can assist in the analysis of texts (section 5.3.2.3). Its use in interlingua-based systems goes somewhat further. In this case the intention is not just to resolve ambiguities in the source text but to derive language-independent representations. Returning to the *wear* example (section 6.1 above), the knowledge base of a true interlingua ought to be capable of distinguishing the sixteen (or more) different aspects of wearing and putting on clothes. It would do so, presumably, on the basis of knowledge about different types of apparel (hats, coats, socks, etc.), about social, cultural, ethnic differences (ceremonial dress, working clothes, children's wear, tropical clothing, Arctic clothing, rainwear, etc.), and about the activities involved with them (making, buying, wearing, removing, washing, repairing, etc.)

The practical complexity of the task inevitably restricts the approach to highly restricted domains with relatively narrow contexts and applications. It is also true to say that knowledge-based MT systems have so far concentrated on the disambiguation of source texts and have not tackled the derivation of universal language-independent representations capable of providing information for generation in any other language. One MT project which is exploring the possibilities is described in section 18.1 below (the KBMT project at Carnegie Mellon University).

6.9 Example-based methods

Developments in computer technology, which are providing faster access to larger memories and data stores, are encouraging the investigation of methods based on access to huge corpora of texts in source and target languages. The basic argument is that translation is often a matter of finding or recalling analogous examples, discovering or remembering how a particular source language expression or something similar to it has been translated before.

Proposals for **example-based** methods are put forward generally as alternatives to knowledge-based approaches and as supplementary aids to the traditional rule-based methods of analysis, transfer and generation. The linguistic databank of 'examples' would be derived from a structural analysis of a large corpus of source texts and their translations in a target language produced by human translators. The result is bilingual sets of phrases; for example, English phrases containing *field* and their French equivalents in the database (Table 6.1).

the main fields	*les principaux domaines*
the following fields	*les domaines suivants*
these two fields	*ces deux domaines*
the specialized fields	*les domaines spécialisés*
the para-medical fields	*activités paramédicales*
the magnetic fields	*les champs magnétiques*
the coal fields	*les bassins-houiliers*
the coal fields	*les bassins*
the corn fields	*les champs de blé*

Table 6.1 Database of aligned examples

Whether *field* is to be translated as *domaine*, *champ* or *bassin* is determined by the frequency of phrases most similar in context to the given example. Exact matches will obviously cause few problems, but they will be rare; for instance, in the list above there is no match for *gold field*. The identification of 'similarity' depends on some measure of distance of meaning; in current proposals this is based on the classification of lexical items in semantic hierarchies, e.g. *field* might be listed under the heading 'range' along with *scope*, *sphere* and *arena*; and under 'enclosure' with *compound*, *yard* and *paddock*; and so forth. In the case of *gold field*, the hierarchy would indicate a smaller distance from *gold* to *coal* than from *gold* to *corn* and hence a higher probability for the translation of *field* as *bassin* than as *champ*.

The example-based approach is particularly attractive for the translation of complex noun phrases. Generally, Japanese noun phrases of the form 'N_1 *no* N_2' correspond to English noun phrases of the form 'N_2 *of* N_1'; but there are many exceptions: it is more idiomatic to say *application fee for the conference* than *application fee of the conference*, *conference in Tokyo* rather than *conference of Tokyo*, and literal translations from Japanese such as **holiday of a week*, **reservation of hotel* or **hotels of three* are not permitted in place of *week's holiday*, *hotel reservation* and *three hotels*. There are similar difficult areas with other language pairs: consider the multiplicity of translations into English of the French *de*. The availability of analogous examples and probabilistic measures of distance could enable such problematic areas of lexical transfer to be handled more satisfactorily than attempts to devise rules based on grammatical categories and structures and the presence or absence of case roles, semantic features and the like. (We describe a proposed example-based MT system in more detail in section 18.2.)

Example-based techniques can also be used for translating sentences which are structurally similar to previously translated sentences. Again a similarity measure is required, though in this case it will be based on distribution of key elements such as grammatical words, or it may measure the similarity of certain sequences of grammatical categories, or a combination of these. For instance, the sample sentence in (47a) might be matched against any text consisting of a sequence such as that in (47b), where X may be any noun or adjective-plus-noun, and Y may be

any noun phrase, and its translation a copy of the translation of (47a), with the non-matching parts translated separately.

(47a) Remove the bulb and replace it with a new one.

(47b) Remove the X and replace it with Y.

This approach could be likened to the use by a (linguistically sophisticated) tourist of a phrasebook, where new utterances can be built up by combining the appropriate elements of the phrases given. (For an experimental system based on this idea, see section 18.5.)

The example-based method should be distinguished from **corpus-based** approaches to MT. In the latter, systems (of whatever type) are developed for translating a subset of texts in a particular corpus and then applied to the translation of other texts in the same corpus. Example-based methods are not necessarily restricted either to particular corpora or to particular sublanguages (section 8.4). Initially, given the effort involved in building the databases, the method will almost certainly be applied within specific domains. However, as far as general vocabulary is concerned, examples taken from one corpus may be just as effective for translating texts in another subject area.

Although it is a natural assumption that example-based methods work best with structured sets of bilingual texts, the experiments at IBM mentioned earlier (section 6.3) show that correspondences of units in source and target texts can also be established by statistical means alone. However, to what extent this extreme position proves valid has yet to be demonstrated. (For more on this system see section 18.3.)

Finally, it should be evident that the example-based approach can be integrated in any of the basic models: direct, transfer, and interlingua. Unlike the knowledge-based method, which is dependent on semantic analysis to a high degree of abstraction, it can operate with relatively shallow analyses of surface structures and relatively simple lexical transfer procedures, i.e. it can be applied to any level of transfer: from morphological to interlingual.

6.10 Summary: comparison of transfer-based and interlingua-based systems

This chapter has illustrated the principal problems of converting lexical and structural information from one language into another for each of the three basic MT designs: direct, transfer-based and interlingua-based. At one extreme, the 'shallow' analyses of direct systems impose heavy burdens on bilingual lexicons and inevitably result in *ad hoc* treatments of structural changes. At the other extreme, the 'conceptual meaning' representations required for interlingua-based systems demand a complexity of semantic analysis beyond the limitations of current linguistic theory. It is generally agreed that transfer-based approaches are at present the best foundations for advances in MT. As we have seen, there are a number of options available: different levels of lexical and structural representation, and different potentials for interlingual elements.

In practical terms, the flexibility of the transfer-based approach overcomes one apparent drawback in comparison with interlingua systems, namely the cost of adding a new language to the system (section 4.2). Although in a multilingual system the transfer methodology requires the construction of a large number of transfer modules, the analysis modules are normally (as we have seen) simpler than interlingual representations, and they can be varied in depth and abstractness according to the languages concerned.

In certain respects, the differences between transfer and interlingua systems are relative. In practice, many systems which are described as 'interlingual' deal only with a single pair of languages, so they differ little in 'depth' of abstractness and universality from many transfer-based systems. But the main point which needs to be stressed is that whatever the framework the same problems of lexical and structural transfer are present: whether in an interlingua or a transfer-based system, the problems of knowing how to translate a certain construction (passive, say) or how to deal with a transfer ambiguity (*wall* into German) remain the same.

Within the transfer-based framework there are also many possible variants. The diagrammatic presentation in Figure 6.1 above is evidently a simplification. Transfer does not necessarily cut across the pyramid horizontally, i.e. between representations of equivalent abstraction. It can be imagined in some cases as cutting across diagonally, representing the respective amount of work done in analysis, transfer and generation, as shown in Figure 6.2. Also, where we have transfer between closely related language pairs having fairly similar structures, a relatively simple transfer can be achieved with a shallower degree of analysis taking advantage of the similarities, i.e. the base of the pyramid would be narrower to start with.

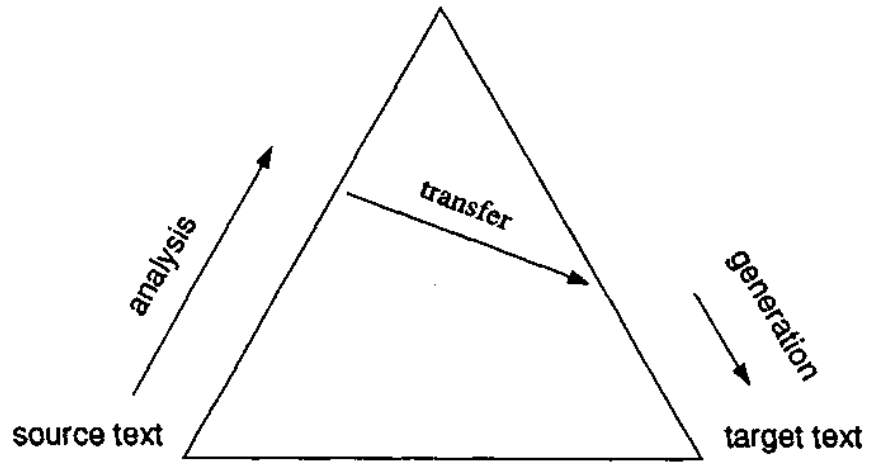


Figure 6.2 Modified pyramid

From a theoretical point of view there is one major advantage in the transfer-based approach. Although translation must involve monolingual linguistic analysis,

it is concerned primarily with questions of **contrastive linguistic analysis**, namely the comparison of the devices that two (or more) languages employ to convey similar meanings. In the interlingua approach, all the contrastive linguistics is hidden in the monolingual analysis and generation, so there can be no separation of the linguistic theories driving the source language analysis, the bilingual transfer linguistics, and the target language generation. Any element of contrastive linguistics is implicit in the structure of the interlingua. With an interlingua system it is difficult to focus specifically on the contrastive aspects of translation, because the interlingua forces the linguist to abstract away from that aspect of translation. In a transfer system however, it is precisely in the transfer module that those purely contrastive differences are captured and dealt with, by whichever method or technique that is most appropriate: lexical substitution, structural change, universal items or features, reference to a knowledge base, statistical probabilities — in fact any of the options described in this chapter.

6.11 Further reading

General discussions of the transfer vs. interlingua problem are to be found in Krauwer and des Tombe (1984), Tucker (1987:22–26), Somers (1987b), Tsujii (1986, 1988) and Boitet (1988).

Examples of lexical translation difficulties are taken from many sources: indeed some of the examples are so widely cited as to have become clichés. The issue of the Eskimo word for 'snow' is discussed by Martin (1986) and Pullum (1989).

Similarly, structural translation difficulties are widely discussed, with the '*like/gern*' problem taken as the standard example.

The 'pyramid' diagram in Figure 6.1 is now a universal symbol, but was first used by the Grenoble team, and probably first appeared (in a slightly different form) in Vauquois (1968).

The 'animate subject constraint' is discussed in Nishida *et al.* (1980). Nominalizations are the subject of a paper by Somers *et al.* (1988).

Systems which use unification-based linguistic techniques and functional representations typify the 'structural interlingua with lexical transfer' approach and include the MiMo2 system (van Noord *et al.* 1990) and basic research work by Rohrer (1986), Kaplan *et al.* (1989), Zajac (1989, 1990) and Sadler *et al.* (1990). The issue is also discussed in Luckhardt (1987).

There have been several notable attempts at interlingua-based MT (see Hutchins 1986:Ch.10). Besides those mentioned in the text, contemporary approaches include those of the Japanese commercial systems of Fujitsu (Uchida 1988, 1989) and NEC (Harada 1986).

Using formulae of propositional logic is an approach currently being investigated by Hobbs and Kameyama (1990). References for the DLT system, which uses Esperanto as an interlingua, are given in Chapter 17. Carnegie Mellon's Knowledge-Based MT is discussed in section 18.1.

Experiments in Example-Based MT are described by Sato and Nagao (1990) (who call it 'Memory-based MT') and Sumita *et al.* (1990). The EBMT system

proposed by the DLT group is described in more detail in section 18.2. The statistics-based approach of the IBM researchers is described in section 18.3.